**Design Document – Financial RAG Assignment**

**Chunking Strategy**

* Used **paragraph + sentence splitting** with token count control (~800 tokens target, max 1000).
* Ensures each chunk is small enough for embedding while keeping financial tables intact.
* Long paragraphs (>1000 tokens) are further split by words with ~50–100 word overlap to avoid context loss.

**Embedding Model Choice**

* **Model**: sentence-transformers/all-MiniLM-L6-v2.
* **Reason**:
  + Lightweight and fast for local FAISS indexing.
  + Good semantic performance for financial Q&A tasks.
  + Open-source, no external API dependency → reproducible.

**Agent / Query Decomposition**

* **Basic queries**:  
  Pattern-matched when company, metric, and year are explicitly present (e.g., *“Microsoft revenue 2023”*).
* **Comparative / multi-company**:  
  Decomposed into sub-queries for each (company, year, metric) pair, answers aggregated.
* **AI / risk-based queries**:  
  Routed to text-based retrieval (no numeric extraction), returning relevant excerpts.
* **Fallback mode**:  
  If no match, run across all companies/years to ensure coverage.

**Implementation Flow**

1. **Acquire filings** → Download 10-Ks from SEC EDGAR by CIK & year.
2. **Parse text** → Convert HTML/PDF to plain text.
3. **Chunk** → Split into ~800 token windows.
4. **Embed + Index** → Use FAISS for similarity search.
5. **Decompose query** → Classify into basic, comparative, AI, or fallback.
6. **Retrieve + Extract** → Pull top-k chunks and apply regex/numeric extraction.
7. **Synthesize answer** → Combine sub-query results with reasoning.

**Challenges & Key Decisions**

* **Noisy filings**: SEC docs have tables, footnotes, and inconsistent formatting → required flexible regex.
* **Numeric extraction**: Difficult to distinguish *revenue* vs. *expense* vs. *margin*; heuristic rules applied.
* **Synonyms**: Needed to expand metric keywords (e.g., *“operating margin” ≈ “operating income”*).
* **Trade-off**: Prioritized speed & transparency over perfect accuracy — aimed for a working prototype, not production-grade financial analytics.